

APPARATUS AND METHOD FOR HAIRPINNING DATA
PACKETS IN AN ETHERNET MAC CHIP

Inventor(s):

Jack C. Wybenga
2129 Stone Creek
Plano
Collin County
Texas 75075
United States citizen

Patricia K. Sturm
4424 Rushing Road
Dallas
Collin County
Texas 75287
United States citizen

Steven E. Tharp
405 Rivercove Drive
Garland
Dallas County
Texas 75044
United States citizen

Youngil Kim
3005 Huffman Drive
Plano
Collin County
Texas 75025
Citizen of Republic of Korea

Assignee:

SAMSUNG ELECTRONICS Co., LTD.
416, Maetan-dong, Paldal-gu
Suwon-city, Kyungki-do
Republic of Korea

William A. Munck
John T. Mockler
Davis Munck, P.C.
Three Galleria Tower
13155 Noel Road, Suite 900
Dallas, Texas 75240
(972) 628-3600

APPARATUS AND METHOD FOR HAIRPINNING DATA
PACKETS IN AN ETHERNET MAC CHIP

5 TECHNICAL FIELD OF THE INVENTION

[001] The present invention relates to massively parallel routers and, more specifically, to a massively parallel, distributed architecture router that transfers data between an ingress processor and an egress processor in the same routing node.

10

BACKGROUND OF THE INVENTION

[002] There has been explosive growth in Internet traffic due to the increased number of Internet users, various service demands from those users, the implementation of new services, such as voice-over-IP (VoIP) or streaming applications, and the development of mobile Internet. Conventional routers, which act as relaying nodes connected to sub-networks or other routers, have accomplished their roles well, in situations in which the time required to process packets, determine their destinations, and forward the packets to the destinations is usually smaller than the transmission time on network paths. More recently, however, the packet transmission capabilities of high-bandwidth network paths

15

20

and the increases in Internet traffic have combined to outpace the processing capacities of conventional routers.

[003] This has led to the development of a new generation of massively parallel, distributed architecture routers. A distributed architecture router typically comprises a large number of routing nodes that are coupled to a high-bandwidth crossbar switch via a plurality of switch fabric modules. Each routing node has its own routing (or forwarding) table for forwarding data packets via other routing nodes to a destination address.

[004] In a conventional 1 Gigabit Ethernet (GbE) router, each routing node uses a network processor (NP) that performs routing functions for both incoming (ingress) and outgoing (egress) data packets. A 1 GbE network processor typically uses micro-engines that forward data packets, both within the same routing node as well as to other routing nodes via the switch fabrics.

[005] The single network processor arrangement is particularly advantageous for "hairpinning" a received data packet. If a routing node receives a data packet from an external source device, and both the source device and the destination device are coupled to the routing node via the physical media device (PMD) modules of the routing node, there is no need to transfer the data packet

through the switch fabrics and/or cross-bar switch associated with the routing node. Instead, the micro-engines of the network processor simply transmit the received data packet back out through the PMD modules (i.e., like a "hairpin" turn) without using the
5 switch fabrics.

[006] However, the single network processor approach cannot keep up with the 10 Gbps data rate of newer 10 Gigabit Ethernet (GbE) routers. As a result, 10 GbE routers use two network processors, an ingress processor for routing incoming data packets
10 received from a PMD module and an egress processor for routing outgoing data packets received from the switch fabric and/or cross-bar switch. Use of two network processors allows higher data rates, but also requires a different approach to hair-pinning data packets, since the micro-engines in the two processors are not
15 handling both the incoming and the outgoing data packets.

[007] An alternative way to hairpin data packets is to let the control plane processors forward the data packets between the network processors within the routing node. However, using a single processor to handle data packets and to perform control
20 plane functions can only be used with very low speed routers. In fact, using a central processor for data plane and control plane

functionality is even worse in terms of the data rates supported than a single network processor in a 1 GbE router configuration.

[008] A new data plane interface could be added between the micro-engines in the egress and ingress network processors to allow the micro-engines of the two network processors to forward packets directly to the other network processor. However, using a new high-speed data interface between the micro-engines of the two network processors is impractical, as there is no convenient place to connect the new interface to the micro-engines. In essence, this approach requires added hardware and complexity, which leads to circuit board problems associated with higher speed interfaces.

[009] Finally, the switch fabric modules could be used to allow loop-backs of data packets during normal operations. However, conventional switch fabrics do not provide this feature. Modification of the switch fabric is possible, but has the disadvantage of requiring a switch module even with a single routing node system. In addition, it has the disadvantage of shipping all of the data back and forth over the optical links, adding to the delays and optical traffic.

[010] Therefore, there is a need in the art for an improved high-speed router. In particular, there is a need for a high

bandwidth mechanism for routing (or hairpinning) data packets between the ingress and egress network processors within the same routing node.

SUMMARY OF THE INVENTION

[011] The present invention comprises a distributed router having multiple routing nodes interconnected by switch fabric modules and an optional backbone crossbar switch. To handle the high-speed interfaces of 10 Gbps routing nodes, the present invention uses separate network processors to handle ingress and egress packets. For packets that are routed between two ports of the same routing node, the present invention uses a modified Media Access Control (MAC) chip that allows data packets to flow between the two network processors, in addition to the normal flow of packets to the switch fabric. The MAC device is modified to provide additional internal interfaces to send packets between the two network processors. These interfaces are internal to the MAC chip. This reduces board layout problems associated with an added high-speed interface.

[012] Modifying the MAC device in this manner ensures that the functionality for hairpinning data packets stays in the routing node and does not involve the switch fabrics. The present invention can therefore support hairpinning at the full 10 Gbps data rate and allows a single routing node to operate without using the switch fabric. Advantageously, the present invention uses

existing interfaces and parts and reduces complexity relative to the alternative of supporting an additional interface.

[013] To address the above-discussed deficiencies of the prior art, it is a primary object of the present invention to provide a
5 router for interconnecting N interfacing peripheral devices. According to an advantageous embodiment, the router comprises a plurality of routing nodes coupled to one another via switching circuitry. A first one of the plurality of routing nodes comprises: 1) at least one physical medium device (PMD) module
10 capable of transmitting data packets to and receiving data packets from selected ones of the N interfacing peripheral devices; 2) an ingress processor capable of receiving incoming data packets from the at least one PMD module; 3) an egress processor capable of transmitting data packets to the at least one PMD module; and 4) a
15 medium access control (MAC) processor capable of forwarding data packets from the ingress processor to the switching circuitry and forwarding data packets from the switching circuitry to the egress processor, wherein the MAC processor is capable of determining whether a first data packet received from the ingress processor is
20 directed to the egress processor and, in response to the determination, transferring the first data packet directly to the

egress processor without forwarding the first data packet through the switching circuitry.

[014] According to one embodiment of the present invention, the MAC processor determines whether the first data packet is directed
5 to the egress processor by an address of the first data packet.

[015] According to another embodiment of the present invention, the MAC processor determines that the first data packet is directed to the egress processor if the address of the first data packet matches an address of the MAC processor.

10 [016] According to still another embodiment of the present invention, the MAC processor forwards the first data packet to the switching circuitry if the address of the first data packet does not match the address of the MAC processor.

[017] According to yet another embodiment of the present
15 invention, the MAC processor comprises a field programmable gate array.

[018] According to a further embodiment of the present invention, the MAC processor comprises a transmit state machine capable of determining whether the first data packet is directed to
20 the egress processor.

[019] According to a still further embodiment of the present invention, the MAC processor comprises a receive state machine coupled to the transmit state machine capable of receiving the first data packet from the transmit state machine if the transmit
5 state machine determines the first data packet is directed to the egress processor.

[020] According to a yet further embodiment of the present invention, the receive state machine forwards the first data packet from the transmit state machine to the egress processor.

10 [021] Before undertaking the DETAILED DESCRIPTION OF THE INVENTION below, it may be advantageous to set forth definitions of certain words and phrases used throughout this patent document: the terms "include" and "comprise," as well as derivatives thereof, mean inclusion without limitation; the term "or," is inclusive,
15 meaning and/or; the phrases "associated with" and "associated therewith," as well as derivatives thereof, may mean to include, be included within, interconnect with, contain, be contained within, connect to or with, couple to or with, be communicable with, cooperate with, interleave, juxtapose, be proximate to, be bound to
20 or with, have, have a property of, or the like; and the term "controller" means any device, system or part thereof that controls

at least one operation, such a device may be implemented in hardware, firmware or software, or some combination of at least two of the same. It should be noted that the functionality associated with any particular controller may be centralized or distributed, 5 whether locally or remotely. Definitions for certain words and phrases are provided throughout this patent document, those of ordinary skill in the art should understand that in many, if not most instances, such definitions apply to prior, as well as future uses of such defined words and phrases.

BRIEF DESCRIPTION OF THE DRAWINGS

[022] For a more complete understanding of the present invention and its advantages, reference is now made to the following description taken in conjunction with the accompanying
5 drawings, in which like reference numerals represent like parts:

[023] FIGURE 1 illustrates an exemplary distributed architecture router that implements a MAC device that sends data packets between two network processors according to the principles of the present invention;

10 [024] FIGURE 2 illustrates the hairpinning of data packets in a routing node containing a single network processor;

[025] FIGURE 3 illustrates the hairpinning of data packets in a routing node containing two network processors according to one embodiment of the present invention;

15 [026] FIGURE 4 illustrates in greater detail the medium access control (MAC) processor in the routing node in FIGURE 3 according to one embodiment of the present invention; and

[027] FIGURE 5 is a flow diagram illustrating the hair-pinning of a data packet in a routing node according to the principles of
20 the present invention.

DETAILED DESCRIPTION OF THE INVENTION

[028] FIGURES 1 through 5, discussed below, and the various embodiments used to describe the principles of the present invention in this patent document are by way of illustration only and should not be construed in any way to limit the scope of the invention. Those skilled in the art will understand that the principles of the present invention may be implemented in any suitably arranged distributed router.

[029] FIGURE 1 illustrates exemplary distributed architecture router 100, which implements a MAC device that sends data packets between two network processors according to the principles of the present invention. According to the exemplary embodiment, router 100 comprises a plurality of rack-mounted shelves, including exemplary shelves 110, 120 and 130, that are coupled via crossbar switch 140. In an advantageous embodiment, crossbar switch is a 10 Gigabit Ethernet (10 GbE) crossbar operating at 10 gigabits per second (Gbps).

[030] Exemplary shelf 110 comprises a pair of redundant switch modules, namely primary switch module (SWM) 114 and secondary switch module (SWM) 116, a plurality of route processing modules 112, including exemplary route processing module (RPM) 112a, RPM

112b, and RPM 112c, and a plurality of physical media device (PMD) modules 111, including exemplary PMD modules 111a, 111b, 111c, 111d, 111e, and 111f. Each PMD module 111 transmits and receives data packets via a plurality of data lines connected to each PMD
5 module 111.

[031] Similarly, shelf 120 comprises a pair of redundant switch modules, namely primary SWM 124 and secondary SWM 126, a plurality of route processing modules 122, including RPM 122a, RPM 122b, and RPM 122c, and a plurality of physical media device (PMD) modules
10 121, including PMD modules 121a-121f. Each PMD module 121 transmits and receives data packets via a plurality of data lines connected to each PMD module 121.

[032] Additionally, shelf 130 comprises redundant switch modules, namely primary SWM 134 and secondary SWM 136, a plurality
15 of route processing modules 132, including RPM 132a, RPM 132b, and RPM 132c, and a plurality of physical media device (PMD) modules 131, including PMD modules 131a-131f. Each PMD module 131 transmits and receives data packets via a plurality of data lines connected to each PMD module 131.

20 [033] Router 100 provides scalability and high-performance using up to M independent routing nodes (RN). Each routing node

comprises a route processing module (RPM) and at least one physical medium device (PMD) module. Each route processing module buffers incoming Internet protocol (IP) frames and MPLS frames from subnets or adjacent routers. Additionally, each RPM classifies requested
5 services, looks up destination addresses from frame headers or data fields, and forwards frames to the outbound RPM. Moreover, each RPM also maintains an internal routing table determined from routing protocol messages and provisioned static routes and computes the optimal data paths from the routing table. Each RPM
10 processes an incoming frame from one of its PMD modules. According to an advantageous embodiment, each PMD module encapsulates an incoming frame (or cell) from an IP network (or ATM switch) for processing in a route processing module and performs bus conversion functions.

15 [034] Incoming data packets may be forwarded within router 100 in a number of different ways, depending on whether the source and destination ports are associated with the same or different PMD modules, the same or different route processing modules, and the same or different switch modules. Since each RPM is coupled to two
20 redundant switch modules, the redundant switch modules are regarded as the same switch module. Thus, the term "different switch

modules" refers to distinct switch modules located in different ones of shelves 110, 120 and 130.

[035] In a first type of data flow, an incoming data packet may be received on a source port on PMD module 121f and be directed to a destination port on PMD module 131a. In this first case, the source and destination ports are associated with different route processing modules (i.e., RPM 122c and RPM 132a) and different switch modules (i.e., SWM 126 and SWM 134). The data packet must be forwarded from PMD module 121f all the way through crossbar switch 140 in order to reach the destination port on PMD module 131a.

[036] In a second type of data flow, an incoming data packet may be received on a source port on PMD module 121a and be directed to a destination port on PMD module 121c. In this second case, the source and destination ports are associated with different route processing modules (i.e., RPM 122a and RPM 122b), but the same switch module (i.e., SWM 124). The data packet does not need to be forwarded to crossbar switch 140, but still must pass through SWM 124.

[037] In a third type of data flow, an incoming data packet may be received on a source port on PMD module 111c and be directed to

a destination port on PMD module 111d. In this third case, the source and destination ports are associated with different PMD modules, but the same route processing module (i.e., RPM 112b). The data packet must be forwarded to RPM 112b, but does not need to
5 be forwarded to crossbar switch 140 or to switch modules 114 and 116.

[038] Finally, in a fourth type of data flow, an incoming data packet may be received on a source port on PMD module 111a and be directed to a destination port on PMD module 111a. In this fourth
10 case, the source and destination ports are associated with the same PMD module and the same route processing module (i.e., RPM 112a). The data packet still must be forwarded to RPM 112a, but does not need to be forwarded to crossbar switch 140 or to switch modules 114 and 116.

15 [039] The third and fourth cases described above are ideal situations for hairpinning data packets in the route processing modules, since the data packets do not need to be forwarded to the switch modules or the crossbar switch. FIGURE 2 illustrates the hairpinning of data packets in route processing module 112a, which
20 contains a single network processor (NP). In the exemplary embodiment, RPM 112a operates at 1 gigabit per second (1 Gbps) and

network processor (NP) 210 is an IXP 1200 processor. At this speed, an IXP 1200 processor is capable of hairpinning data packets, as shown by the dotted line, using the internal micro-engines of the IXP 1200. The micro-engines simply transmit the
5 received data packet back through the PMD modules without using the switch modules or the crossbar switch.

[040] However, the configuration in FIGURE 2 is inadequate for the speeds in a 10 Gigabit Ethernet router. FIGURE 3 illustrates the hairpinning of data packets in route processing node 112a,
10 which contains two network processors according to an exemplary embodiment of the present invention. Route processing module (RPM) 112a comprises egress network processor (NP) 310, ingress network processor (NP) 320, and medium access control (MAC) processor 330.

In the exemplary embodiment, RPM 112a operates at 10 gigabit per
15 second (10 Gbps). Ingress network processor (NP) 320 is, for example, an IXP 2800 processor that receives data packets from the PMD modules and forwards the received data packets to MAC processor 330. Egress network processor (NP) 310 is an IXP 2800 processor that receives data packets from MAC processor 330 and forwards the
20 received data packets to the PMD modules. In the exemplary

embodiment, MAC processor 330 may comprise a Xilinx field programmable array (FPGA).

[041] FIGURE 4 illustrates medium access control (MAC) processor 330 in route processing module 112a according to the principles of the present invention. MAC processor 330 comprises output interface (IF) 410, input interface (IF) 420, receive (RX) state machine 430 and transmit (TX) state machine 440. Input IF 420 buffers incoming data packets from the PMD modules and output IF 410 buffers outgoing data packets being sent to the PMD modules.

10 The present invention provides a mechanism by which TX state machine 440 compares the destination address (DA) of the data packet to the MAC address of MAC processor 330. If the destination address is not equal to the MAC address, TX state machine 440 forwards the data packet to the switching module. If the

15 destination address is equal to the MAC address, TX state machine 440 forwards the data packet to the RX state machine 430.

[042] FIGURE 5 depicts flow diagram 500, which illustrates the hair-pinning of a data packet according to the principles of the present invention. Initially, ingress network processor (NP) 320 forwards the received data packet to medium access control (MAC) processor 330 (process step 505). Next, transmit (TX) state

20

machine 440 compares the address of the data packet to the address of MAC processor 330 (process step 510). If the addresses are not equal, TX state machine 440 forwards the data packet to switch module (SWM) 114 (process steps 515 and 520). If the addresses are
5 equal, TX state machine 440 forwards the data packet to RX state machine 430 (process steps 515 and 525). Finally, RX state machine 430 forwards the data packet to egress network processor (NP) 310 (process step 530).

[043] Although the present invention has been described with an
10 exemplary embodiment, various changes and modifications may be suggested to one skilled in the art. It is intended that the present invention encompass such changes and modifications as fall within the scope of the appended claims.